第 36 卷 第 2 期
2007 年 5 月
测 绘 学 报
ACTA GEODAETICA et CARTOGRAPHICA SINICA
Vol. 36, No. 2
May, 2007

文章编号: 1001-1595(2007)02-0210-08 中图分类号: P208 文献标识码: A

基于概率的地图实体匹配方法

童小华[1,2], 邓慓慓[1], 史文中[2]
(1. 同济大学 测量与国土信息工程系, 上海 200092; 2. 香港理工大学 土地测量与地理资讯学系, 香港 九龙)

A Probabilistic Theory-based Matching Method

TONG Xiao-hua[1,2], DENG Su-su[1], SHI Wen-zhong[2]
(1. *Department of Surveying and Geo-informatics, Tongji University, Shanghai* 200092, *China*; 2. *Department of Land Surveying & Geo-informatics, The Hong Kong Polytechnic University, Hong Kong, China*)

Abstract: The conflation of geographic datasets is one of the key technologies in the front research area of spatial data capture and integration in Geographic Information Systems (GIS). Map conflation is a complex process of matching and merging map data. Because various reasons relate to map data discrepancies, a great amount of uncertainties exist during the process. In the first step, selecting appropriate thresholds and handling one-many or many-many matching relationships are two difficulties in feature matching, which predetermines following map merging step. This paper proposed a probabilistic method for feature matching, which fuses a variety of criteria to calculate the matching probability. The feature pair with the highest probability can be determined to be matched. This method avoids selecting thresholds and attempts to resolve one-many and many-many matching relationship.

Key words: map conflation; probabilistic theory; feature matching; multi-indicators fusion

摘 要: 数字地图合并是通过同名实体匹配和合并变换技术, 调整相关地物实体的几何、属性等差异, 实现同一地区不同来源地图数据的集成和融合。其中同名实体匹配是极为重要的第一步, 也是一个存在大量不确定性的过程, 匹配阈值的选取、实体非一对一的匹配关系是匹配中的关键难题, 匹配效果不佳或出现错误匹配直接影响着后续合并结果的正确性。本文提出一种基于概率理论的匹配模型, 该模型融合多种匹配指标, 通过计算实体匹配概率大小来确定匹配实体。该方法避免了匹配指标精确阈值的选取, 并且能够有效地解决匹配中非一对一的情况。

关键词: 数字地图合并; 概率理论; 匹配; 多指标融合

1 引 言

GIS 应用中面临的一个难题是如何将不同来源、不同程度差异的地理信息数据集融合和集成, 生成满足某种要求的新的数据集, 以达到数据复用的目的。地图数据合并技术是解决这一问题的关键技术。其基本原理是在同名实体匹配的基础上, 建立两个或两个以上地图数据库之间的局部坐标转换关系, 从而获得图形和属性数据的融合, 实现同一地区不同来源地图数据库的集成。一般是依据一幅图对另一幅图进行变换, 依据的图称作"参考图", 调整的图称作"调整图"。地图合并主要由两个过程构成: 同名实体的识别或匹配以及合并变换。匹配是地图合并过程的第一步, 如果没有较好的匹配效果作为合并变换的基础, 会影响整个地图合并的结果。

国内外学者文献研究了多种实体匹配的方法。Saalfeld 提出了结合点的几何位置与蜘蛛编码的点匹配方法[1,2], 能识别一对一匹配的情况, 而没有考虑一对多或多对多的匹配情况。Gabay 和 Doytsher 提出利用点距及线段方向夹角作为匹配指标进行线串的匹配[4]。Cobb 等提出了基于知识的非空间属性数据匹配策略[5], 通过计算属性项的相似度值以确定匹配实体, 也未考虑非

收稿日期: 2006-10-13; 修回日期: 2006-12-15
基金项目: 国家自然科学基金项目(40301043); 上海市青年科技启明星计划项目(05QMX1456); 地理空间信息工程国家测绘局重点实验室经费项目(200618)
作者简介: 童小华(1971-), 男, 江西东乡人, 教授, 博士生导师, 主要研究方向为遥感和 GIS 数据不确定性处理与质量控制。
E-mail: tongxhtj@yeah.net

一对一的匹配情况。目前大多数匹配方法是先进行点匹配, 在此基础上再进行线匹配、面匹配。匹配的过程中的一个难题是匹配指标阈值的选取, 若选取过大的阈值, 会增加误匹配的几率, 而阈值过小则会减少匹配点对, 从而影响后续的合并变换操作。另一个不足在于许多方法未考虑非一对一的匹配情况(其产生原因在于由于待合并地图的比例尺不同, 或地图综合的影响, 较小比例尺图中的一个实体可能对应大比例尺图中的多个实体) 。Walter 和 Fritsch 提出了基于概率统计的匹配方法[6], 利用“缓冲区增长”法确定候选匹配集, 再通过区域统计来确定匹配的阈值, 最后使用信息论中的优势函数(Merit Function) 确定匹配结果, 该模型理论严密, 有较好的匹配效果, 但计算过程复杂, 较为费时, 同时匹配阈值对结果也有影响。由此, 本文提出了一种基于概率理论的匹配模型, 综合运用了多个匹配指标, 通过计算实体匹配概率大小来确定匹配实体, 同时避免了精确阈值的选取, 且有效地解决了匹配中非一对一的情况。

2　主要匹配算法

2.1　非概率匹配算法

Saalfeld 最早提出了地图合并的匹配算法, 使用距离、度和蜘蛛编码寻找匹配点, 然后利用已匹配点构成三角网对未匹配点进行变换, 不断迭代进行点匹配和坐标变换直至不再出现新的匹配点为止。该方法的前提为两幅图是拓扑同构的, 即只能解决一对一的匹配。

Gabay 和 Doytsher 首先使用点距离信息确定匹配点, 分为匹配顶点及匹配结点, 之后进行多义线匹配, 对具有匹配结点的待匹配多义线建立缓冲区, 缓冲区大小根据地图的点位精度而定, 距离阈值计算为

$$\text{距离阈值} = k \times \sqrt{m_a^2 + m_b^2} \qquad (1)$$

式中, k 是一个常数, 一般取 3, m_a 是图层 A 的点位精度, m_b 是图层 B 的点位精度。从匹配结点开始, 对参考图中的多义线逐线段检验是否落在缓冲区内, 以及两线段方向夹角是否小于某一阈值, 方向夹角阈值的计算式为

$$\text{方向夹角阈值} = k \times \sqrt{2\times (m_a/d_a)^2 + 2\times (m_b/d_b)^2} \qquad (2)$$

式中, k, m_a, m_b 的定义同式(1), d_a 表示 A 图中线段的长度, d_b 表示 B 图中线段的长度。如果一条线段落入对应线段的缓冲区内且夹角小于阈值, 则两线段匹配, 若两条多义线的每一个对应线段均匹配, 则两条多义线为匹配实体; 如果仅是多义线的部分线段匹配, 则从匹配线段的端点开始分两个方向匹配, 直到线的结点或无匹配线段为止。

2.2　基于概率统计的匹配方法

基于概率统计理论, 文献[6]首先利用缓冲区增长法确定待匹配线串的候选匹配集, 选取一片已经实现了实体匹配的区域进行统计, 统计所有对应匹配线实体各项指标的差值, 根据这些差值的分布来确定指标阈值, 其中阈值的选取应使得至少 90% 的线实体在这个范围内, 再根据阈值进一步缩小候选匹配集, 然后计算实体间的相关信息(Mutual Information)

$$I_k(a_i; b_j) = \log_2 \frac{P_k(a_i \mid b_j)}{P_k(a_i)} \qquad (3)$$

式中, a_i 是图层 A 中某一实体, b_j 是图层 B 中某一实体, $I_k(a_i; b_j)$ 是 a_i、b_j 两个实体的第 k 个匹配指标的相关信息值, $P_k(a_i)$ 为实体 a_i 的第 k 个匹配指标取某一个值时的概率, 不同的指标有不同的概率计算方法, 如距离指标, 若将整幅图划分为 $n \times m$ 块区域, 则 a_i 的坐标(x_i, y_i)落入某一区域的概率为 $P(a_i) = \frac{1}{n \times m}$, $P_k(a_i \mid b_j)$为 a_i 与 b_j 匹配的条件概率, 由 a_i 与 b_j 的概率按照条件概率的公式计算而得, a_i 与 b_j 匹配的相关信息是所有指标的相关信息的总和。最终通过树搜索的方法得到一个匹配实体集, 使得匹配集中所有线实体的总相关信息(所有实体的相关信息之和) 最大。

为了解决非一对一的情况, 文献[11]依据模糊集理论, 提出了通过计算可信度值寻找匹配实体的方法。假定两个数据集为 $A = \{a_1, a_2, \cdots, a_m\}$ 和 $B = \{b_1, b_2, \cdots, b_n\}$, 对于 A 中某个实体 a_i, 计算 a_i 选择 B 中某个实体 b_j 作为匹配实体的概率为

$$P_{a_i}(b_j) = \frac{\text{distance}(a_i, b_j)^{-\alpha}}{\sum_{k=1}^{n} \text{distance}(a_i, b_k)^{-\alpha}} \qquad (4)$$

式中, α 是退化因子, 决定了 b_j 与 a_i 距离增加时其匹配概率减少的速度, n 是 B 中实体的个数, $\text{distance}(a_i, b_j)$ 是 b_j 与 a_i 的距离, 再反过来计

算 B 中实体 b_j 选择 A 中实体 a_i 作为匹配实体的概率 $P_{b_j}(a_i)$, 最后计算 a_i 与 b_j 匹配的可信度值

$$\text{confidence}(\{a_i, b_j\}) = \sqrt{P_{a_i}(b_j) \cdot P_{b_j}(a_i)} \quad (5)$$

指定一个可信度值的阈值, 若上式计算的可信度值在阈值范围之内则实体对互为匹配实体。这种方法计算较为简单, 为了解决非一对一的情况而设定了阈值, 但如果阈值设定的不恰当, 则会影响匹配的效果。

3 基于概率理论的多指标融合的匹配算法

3.1 匹配算法使用的指标

匹配算法中常用的指标有空间信息指标和非空间属性信息指标, 本文着重探讨利用空间信息指标进行匹配。基于空间信息的指标有以下几种:

①距离指标。同名实体在距离上应较为接近。点实体一般使用欧式距离; 线实体距离的计算较困难, 主要有 Hausdorff 距离、L_2 距离[2] (两条线上相应点的距离累计之和)、基于中间面积法的线实体之间距离[7] (利用两条线实体首尾点围成的面积与线长度的比值算得) 等算法; 面实体的情况更为复杂, 其距离确定可以通过计算重心点间的距离, 也可以将其边界视为线实体, 利用上述线实体距离相似度衡量。② 形状指标。对于线实体, 在人工智能与模式识别领域主要有 Freeman 编码法、函数描述法等。Saalfeld 在 L_2 距离的基础上进行改进, 得到了线实体形状的测度。张桥平提出了一种基于方向变化角的线实体形状相似度的计算方法[7]。面实体形状相似度的计算方法有多种, Wentz 提出了面的紧致度(面积周长比)、边界的描述和面的构成成分来定义面实体的形状[9]。Foley 使用了类似于线实体形状度量的方法, 即通过计算边界线间的面积以求得实体间的形状相似度[10]。③ 方向指标。线实体方向相似度可以通过计算线串首尾结点连线的方向角来表示, 面实体方向相似度的计算一般通过计算面实体的最小外包矩形(MBR), 通过比较 MBR 的对角线方向以确定方向相似度。④实体的结构。点实体的结构可由连接到该点的线实体数(度) 及连接到该点的线实体的方向描述。Saalfeld 提出了一个称为“蜘蛛编码”(Spider Code) 的 8 字节二进制结点结构编码方案用于匹配[1,2]。将一个结点可能的连线方向角分为 8 个连续不相交的角度区域, 并用一个 8 位长的编码来表示结点的结构特征。⑤拓扑信息。一幅图中的未匹配点(线) 与周围匹配点(线) 之间的拓扑关系应与另一幅图中同名点(线) 与该匹配点(线) 间的拓扑关系对应相同。面实体的拓扑相似度可由两个面的重叠面积衡量, 文献[12] 使用成分关联区域来量度拓扑关系。

利用以上某种指标匹配得到的同名点(线) 对并不一定能满足其他指标的条件, 两条具有相似形状的线串并不一定就是同名线串对, 而两个同名点也不一定具有最近的距离。在不能满足所有条件的情况下就需要协调各匹配指标。一般的匹配方法是顺序使用各匹配指标, 例如先计算距离相似度, 排除一些不可能的点, 再计算形状相似度, 进一步排除点。这种匹配策略要确定各项指标的阈值, 而且使用指标的顺序不同, 得到的匹配结果可能不同。因此应整体使用这些指标, 计算出一个总相似度值以确定匹配结果, 并且要考虑非一对一的匹配情况。

3.2 多指标融合的基于概率理论的匹配算法

本文在文献[11] 基于概率的匹配算法基础上进行了扩展, 提出了一种广义的匹配算法, 文献[11] 的方法只利用了一种空间信息指标——距离, 而本文提出的模型融合了多种信息指标, 并且解决了一对多的匹配情况。一对多匹配的存在是客观合理的, 大多是由于地图综合的影响, 将若干个实体综合成为一个实体, 在匹配过程中应该能够识别出这种情况。

设同一地区两个不同图层的数据集分别为 $A = \{a_1, a_2, \cdots, a_m\}$ 和 $B = \{b_1, b_2, \cdots, b_n\}$, 两数据集的实体数目不一定相同, 即 m 可能不等于 n。根据图中实体的特性, 选择若干种空间信息匹配指标, 选择的指标种类越多, 越能增强辨别实体差异的能力。图 A 实体与图 B 实体匹配概率的计算公式为

$$P_{a_i, b_j} = \sum_{l=1}^{r} P_{a_i, b_j}(l) \cdot P_l \quad (6)$$

$$P_{a_i, b_j}(l) = \frac{d(a_i, b_j)^{-\alpha}}{\sum_{k=1}^{s} d(a_i, b_k)^{-\alpha}} \quad (7)$$

式中, r 是所有匹配指标的个数, $P_{a_i, b_j}(l)$ 是针对第 l 个指标时实体 a_i 与 b_j 的匹配概率, P_l 为第 l 个指标的权重, P_{a_i, b_j} 为所有指标概率的加权平均即实体 a_i 与 b_j 匹配的总概率。$d(a_i, b_j)$ 是 a_i 的

指标值与 b_j 的指标值差值的绝对值, 若两指标值精确相等, $d(a_i, b_j)$ 值为 0, $d(a_i, b_j)^{-1}$ 应为无穷大, 则将 $d(a_i, b_j)^{-1}$ 定为较大的值以扩大该点匹配的概率, s 是 a_i 的候选匹配集的实体个数, 在计算 a_i 与候选匹配集中某个实体 b_j 的匹配概率前, 应先计算 a_i 与每一个候选匹配实体的指标差值, α 为退化因子, 本文取值为 1。最终选择与 a_i 匹配的实体时应满足

$$P_{\text{match}} = \max(P_{a_i, b_1}, P_{a_i, b_2}, \cdots, P_{a_i, b_s}) \quad (8)$$

式中, P_{match} 为最终匹配点的匹配概率, P_{a_i, b_1}, P_{a_i, b_2}, …, P_{a_i, b_s} 是候选匹配集中所有实体与 a_i 的匹配概率, 也就是最终匹配点的匹配概率应是候选匹配集中所有实体与 a_i 匹配概率的最大值。

本文算法采取先进行点匹配再进行线匹配或面匹配, 并且需要进行双向匹配, 即先确定 A 在 B 中的匹配实体(称为正向匹配), 再确定 B 在 A 中的匹配实体(称为反向匹配)。匹配步骤为: ①确定 A 中某一实体 a_i 在 B 中的候选匹配集; ②选取匹配指标: 点实体可以选取距离、结构等作为匹配指标, 线实体可以选取距离、形状、方向等作为匹配指标, 面实体则可以选取重心距离、重叠面积等作为匹配指标; ③使用式(6)、式(7)计算候选匹配集中每一实体与 a_i 匹配的概率; ④按照式(8)取概率最大者作为匹配实体, 并将两实体均标识为已匹配; ⑤搜索 B 中未匹配实体, 对未匹配实体 b_j 确定其在 A 中的候选匹配集; ⑥使用式(6)、式(7)计算候选匹配集中每一实体与 b_j 匹配的概率; ⑦按照式(8)取概率最大者作为候选匹配实体, 如果是点匹配, 检查待匹配点 b_j 的相邻点是否与候选匹配点的相邻点匹配, 若至少有一个相邻点匹配, 则确定待匹配点与候选匹配点的匹配关系。

采用双向匹配的策略是为了解决非一对一的匹配情况, 例如 A 中实体 a_1 在 B 中对应的同名匹配实体为 b_1 和 b_2, 在对 A 中实体搜索其匹配实体时, a_1 选择 b_1 作为匹配实体, 对 B 中未匹配实体搜索其匹配实体, b_2 会选择 a_1 作为匹配实体, 最终确定了 a_1 与 b_1 和 b_2 的一对二的匹配关系。确定候选匹配集的方法详见匹配策略部分。

在使用式(6)计算总概率时, 通过给每一种指标赋一个权值, 然后计算指标的加权平均值作为实体的总匹配概率, 以此来加大匹配实体与非匹配实体间的差异。例如两幅图中同名实体结构形状大致相同而差异主要在距离上, 则距离指标可以赋予较大的权重。可以采用经验值法等来定权, 例如由操作人员观察两幅图的差异以确定每一个指标的权。关于匹配指标的定权方法, 将另文介绍。

3.3　匹配策略

根据空间实体的分类, 实体匹配也分为点匹配、线匹配和面匹配, 这三类匹配使用的指标, 步骤各不相同, 但最终匹配实体的确定均基于本文所提出的概率算法。

1. 点匹配

点实体的匹配一般选取距离及所属线或面的信息如线上点的拐角、结构或点线拓扑关系等。为了减少搜索时间, 匹配时首先选取待匹配点在另一幅图中的候选匹配集, 一般指定一个距离范围, 落入此范围的即为候选匹配点。通常使用的匹配指标为距离、度、蜘蛛编码, 在使用蜘蛛编码计算概率时, 若两点的蜘蛛编码完全不同, 就将该点从候选匹配集中剔除, 并且为了减少误匹配的发生, 在双向匹配的第二次匹配时, 要检查待匹配点的相邻点是否与候选匹配点的相邻点匹配。

2. 线匹配

线实体的匹配较为复杂, 一般先匹配线上点, 再匹配线实体。线实体的信息包括长度、方向、线顶点的拐角、线结点的结构、拓扑信息等。本文依据点匹配的结果将线匹配分为四种情况: ①多义线两结点已匹配; ②多义线其中一结点与另一条线的结点匹配; ③多义线其中一结点与另一条线上的顶点匹配; ④多义线两结点均未匹配。线实体的候选匹配集将根据这四种情况进行确定。

第一种情况将另一幅图中两结点为对应匹配点的多义线作为候选匹配实体, 选择距离及形状作为匹配指标, 计算总的匹配概率, 具有最大概率的作为匹配线对。线之间距离的计算可使用 L_2 距离, 形状指标的计算采用基于方向变化角的方法[7]。第二种情况只有一个结点匹配, 将所有具有匹配结点的多义线作为候选匹配实体, 计算总匹配概率时, 选择线与线的距离、线的形状以及方向作为匹配指标。第三种情况为结点与顶点匹配, 多义线被这个顶点分为两个部分, 应分别进行匹配, 过程与第二种情况相同。第四种情况出现的原因较为复杂, 可能是此多义线在另一幅图上消失了, 也有可能是发生了较大的改变, 如河流的改道或道路的改建等, 由于可利用的信息较少, 本文对这种情况不做考虑。

3. 面匹配

与线实体匹配不同, 由于面实体边界的起始点与终止点不明确, 因此使用边界线匹配的方法是不适合的, 应将面实体作为一个整体看待来进行匹配, 首先使用多边形的最小外包矩形(MBR)寻找候选匹配实体, 然后利用面实体的重心距离、重叠面积及其他指标如形状信息等计算匹配概率, 若重叠面积为 0 则从候选匹配集中删除。

4 基于概率理论的匹配算法实验

4.1 实验算例

本文选取了同一地区的地形图与地籍图进行实验。点匹配的实验算例如图 1 所示。细实线与阴影区域分别表示两个图层 A, B, 图层 A 有 21 个点, 图层 B 有 12 个点。选取候选匹配集的距离范围定为 10 m, 匹配使用的指标为距离、度、蜘蛛编码。先后对 A, B 中的点搜索其相应的匹配点, 使用公式(6)、式(7)、式(8)计算匹配概率的同时, 检查待匹配点的相邻点的匹配关系以进一步约束。匹配结果如图 2 所示, 粗虚线表示点对的匹配关系, 共匹配 11 对点(1∶ n 匹配算作 n 对匹配点), 其中 4 对点分别为两个一对二匹配。图 3 中 1 号点与 6′ 号点为一对一匹配, 14 号点与 4′ 号、12′ 号点为一对二的匹配关系。

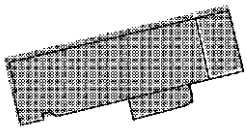

图 1 实验算例一

Fig. 1 Test Example One

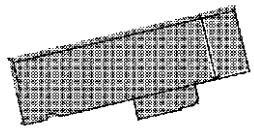

图 2 匹配结果(全局)

Fig. 2 Results of Feature Matching (whole)

表 1 显示了图 3 中的点匹配时的各项指标值之差 $d(a_i, b_j)$ 及概率总和 P_{a_i, b_j} 的计算结果, 其中"蜘蛛编码之差"的计算先比较待匹配点与候选匹配点的 8 位蜘蛛编码[1, 2], 相同位的个数与两点中最小度值的差值即"蜘蛛编码之差"。对于指标值之差为 0 的, $d(a_i, b_j)^{-1}$ 应为无穷大, 在此使 $d(a_i, b_j)^{-1} = 10$。由表中数据可知, 在搜索 A 中 14 号点的匹配点时, 由于 4′ 点的匹配概率最大, 因此选取 4′ 点作为匹配点, 6′ 点与 1 号点匹配, 无反向匹配。在搜索 B 中未匹配实体 12′ 点的匹配实体时, 即可确定 14 点为匹配实体, 从而确定 14 号点与 4′ 号、12′ 号点的一对二的匹配关系。

表 1 点匹配概率计算

Tab. 1 Results of point matching probability

匹配方向	待匹配点	候选匹配点	距离/m	度之差	蜘蛛编码之差	匹配概率和	匹配结果
正向	14	4′	0.82	1	0	2.977	✓
		12′	1.01	1	0	2.493	
		6′	6.03	0	1	1.203	
反向	12′	14	0.82	1	0	2.305	✓
		1	7.06	1	1	0.695	
匹配方向	待匹配点	候选匹配点	距离/m	度之差	蜘蛛编码之差	匹配概率和	匹配结果
正向	1	6′	0.51	1	0	2.025	✓
		4′	5.48	1	1	0.497	
		12′	7.06	1	1	0.558	
反向	无						

线匹配的实验算例如图 4 所示, 细实线表示图层 A, 虚线表示图层 B, 首先匹配线上的点, 方法与点实体匹配相同, 然后根据匹配点的类型分为不同的线匹配情况确定候选匹配实体, 选取 L_2 距离、线的形状及方向作为匹配指标计算匹配概率, 最后根据匹配概率选取匹配线实体。图 5 显示了线实体匹配结果, 图层 A 有 8 条多义线, 图层 B 有 12 条多义线, 短粗虚线表示线上点的匹配关系, "1∶ 1"表示一对一的线匹配关系, "1∶ 2"表示一对二的线匹配关系, 由图中可知一对一匹配共有六对, 一对二匹配共有两对。图 6 与图 7 为两种匹配情况的局部放大图, 箭头表示多义线间的匹配关系, 图 6 中, 图层 A 中的 A_5 与图层 B 中的 B_7 为一对一匹配, 图 7 中, 图层 A 中的 A_1 的顶点与图层 B 中 B_{11}、B_{12} 的端点(结点)匹配, 最终图层 A 中线实体与图层 B 中两个线实体确定了匹配关系。

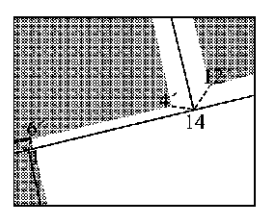

图 3 匹配结果(局部)

Fig. 3 Results of feature matching (part)

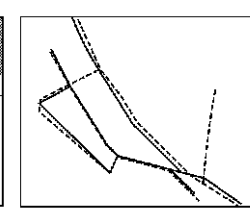

图 4 实验算例二

Fig. 4 Test example two

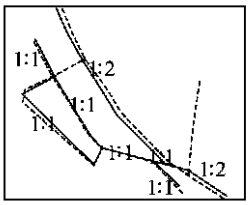

图5　匹配结果(全局)

Fig. 5　Results of feature matching (whole)

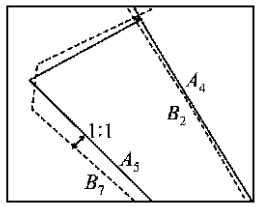

图 6　一对一匹配结果(局部)

Fig. 6　Results of 1∶1 feature matching (part)

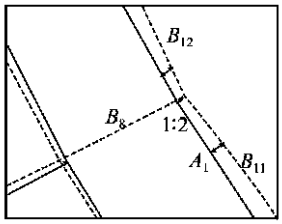

图 7　一对二匹配结果(局部)

Fig. 7　Results of 1∶2 feature matching (part)

表 2 显示了图 6、图 7 中线匹配时的各项指标值之差 $d(a_i, b_j)$ 及概率总和 P_{a_i, b_j} 的计算结果。A_1 与 B_{11}, B_{12}匹配时, 由于是待匹配多义线的顶点与候选匹配线的结点匹配, 而待匹配多义线的结点或其他顶点没有相应的匹配点, 因此最多为一对二的匹配, 不可能有三条多义线同时与其匹配, 正向匹配中 A_1 与概率最大的 B_{11}匹配, 反向匹配中未匹配实体 B_{12}只有一个候选匹配实体 A_1, 因此其同名实体确定为 A_1, 最终 A_1 与 B_{11}, B_{12}为一对二匹配关系。A_5 与 B_7 为一对一匹配, 无反向匹配, 由于两端结点均匹配, 按照匹配策略没有使用方向角指标。

表 2　线匹配概率计算

Tab. 2　Results of Line matching probability

匹配方向	待匹配线	候选匹配线	L_2 距离/m	方向角之差	形状指标之差	匹配概率和	匹配结果
正向	A_1	B_8	349.26	1.52	1.87	0.092	
		B_{11}	49.81	0.01	0.05	1.978	✓
		B_{12}	47.53	0.05	0.14	0.928	
反向	B_{12}	A_1	47.53	0.05	0.14		✓
匹配方向	待匹配线	候选匹配线	L_2 距离/m	方向角之差	形状指标之差	匹配概率和	匹配结果
正向	A_5	B_2	264.68		2.53	0.209 7	
		B_7	31.06		0.29	1.792 1	✓
反向	无						

面匹配的实验算例如图 8 所示, 其中细实线表示图层 A, 阴影部分表示图层 B。匹配面实体时首先根据面实体的最小外包矩形确定候选匹配实体集, 然后选取重心距离、重叠面积作为面实体的匹配指标计算概率进行匹配。匹配结果如图 9 所示, 箭头连接了面实体的重心, 表示不同图层面实体的匹配关系。由图中可知面实体 A_1 与 B_1 为一对一匹配, B_2 与 A_2, A_3 为一对二匹配。

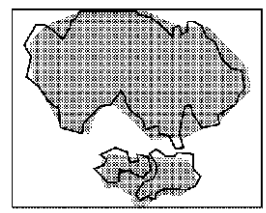

图 8　试验算例三

Fig. 8　Test example three

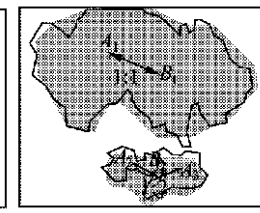

图 9　匹配结果

Fig. 9　Results of feature matching

表 3 为面实体 B_2 与 A_2, A_3 匹配时的各项指标值之差 $d(a_i, b_j)$ 及概率总和 P_{a_i, b_j} 的计算结果, 其中"重叠面积指标之差"为重叠面积与两个多边形面积较小值的差值。通过双向匹配, 确立了 B_2 与 A_2, A_3 一对二的匹配关系。

表 3　面匹配概率计算

Tab. 2　Results of area matching probability

匹配方向	待匹配面	候选匹配面	重叠面积指标之差	重心距离	匹配概率和	匹配结果
正向	B_2	A_2	164 803.7	597.83	0.947	
		A_3	108 077.5	734.54	1.053	✓
反向	A_2	B_2	164 803.7	597.83		✓

4.2　非概率匹配算法与本文算法的比较

本文提出的基于概率理论的匹配算法的主要优点是受阈值的影响较小, 尽量避免了精确阈值的选取, 匹配结果较为稳定。与非概率方法的比较结果如表 4 所示。非概率方法匹配点实体采用距离、度及蜘蛛编码作为匹配指标, 匹配时对各项指标设定阈值, 先使用距离选取候选匹配点, 再使用度、蜘蛛编码进行点的筛选, 在阈值之内的作为匹配点。非概率方法匹配线实体选取 L_2 距离、线实体方向、形状作为匹配指标, 匹配过程与点实体相同。非概率方法匹配面实体选取重心距离与面积重叠度(重叠面积与最小面实体面积的比值)作为匹配指标。

表 4 与非概率方法的比较结果

Tab. 1 Results of comparing with non-probabilistic matching method

试验编号	匹配实体	匹配方法	匹配阈值	匹配实体实际对数	算法匹配实体对数	误匹配个数	漏匹配个数
1	点匹配	非概率	距离≤0.5 m; 度差≤1; 蜘蛛编码相等	66	64	0	2
2	点匹配	非概率	距离≤0.8m; 度差≤1; 蜘蛛编码相等	66	69	3	0
3	点匹配	非概率	距离≤0.5 m; 度差≤1; 蜘蛛编码差≤1	66	65	1	2
4	点匹配	多指标融合概率	距离≤10 m	66	66	0	0
5	线匹配	非概率	L_2 距离≤25 m; 方向差≤0.03(grad); 形状差≤0.03(grad)	10	6	0	4
6	线匹配	非概率	L_2 距离≤25 m; 方向差≤0.05(grad); 形状差≤0.05(grad)	10	7	0	3
7	线匹配	非概率	L_2 距离≤30 m; 方向差≤0.05(grad); 形状差≤0.05(grad)	10	8	0	2
8	线匹配	多指标融合概率	无阈值	10	10	0	0
9	面匹配	非概率方法	重心距离< 20 m; 面积重叠度> 0.8	11	6	0	5
10	面匹配	非概率方法	重心距离< 50 m; 面积重叠度> 0.8	11	10	0	1
11	面匹配	多指标融合概率	无阈值	11	11	0	0

前 4 个试验针对点实体匹配, 分别对本文方法及非概率方法进行了比较。对于非概率方法, 通过变化指标阈值进行检验, 并将两种方法得到的匹配结果与人工判读的实际结果进行比较, 得到误匹配(误匹配即本不该有匹配关系的一对实体确定为匹配实体) 及漏匹配结果。试验 1 距离阈值偏小, 未检验出两个一对二匹配, 试验 2 距离阈值偏大, 虽然检验出两个一对二匹配, 但增加了三个误匹配, 试验 3 放松了蜘蛛编码的约束, 未检验出两个一对二匹配且增加了一个误匹配, 试验 4 采用了本文提出的方法, 只指定了一个较大的选取匹配候选集的距离范围, 最终得到了正确结果, 没有出现误匹配和漏匹配情况。

线匹配试验也对非概率方法及本文的概率方法进行了比较, 可以看出随着不同指标阈值的变化, 非概率方法匹配实体的对数也随之增减, 试验 5 漏匹配了 4 对实体, 试验 6 增加了方向及形状指标的阈值, 漏匹配了 3 对实体, 试验 7 增加了距离指标的阈值, 仍漏匹配 2 对实体, 而多指标融合的概率方法则不需要规定阈值, 仍能得到相对正确的结果。

最后三个实验为面实体匹配, 与点实体、线实体的匹配实验相同, 非概率方法的阈值选取对匹配结果影响较大, 而本文方法可以避免阈值的选取, 并且能够识别非一对一的情况。

5 结 论

地图数据合并过程中, 同名实体的识别或匹配是第一步也是极为重要的一步, 如果没有较好的匹配效果作为合并变换的基础, 会影响整个地图合并的结果。传统的匹配算法在选取指标之后需要确定每一个指标的阈值, 而阈值选择的是否恰当会影响整个匹配的结果。本文提出的基于概率理论的匹配算法, 融合多匹配指标, 充分利用了多种指标信息, 提高了匹配的正确率, 同时尽量避免了阈值的选取, 解决了非一对一的匹配的情况, 得到较好的匹配结果。

参考文献:

[1] SAALFELD A. Automated Map Conflation[D]. Washington D C: University of Maryland, 1993.

[2] SAALFELD A. Conflation: Automated Map Compilation[J]. International Journal of Geographical Information Systems, 1988, 2(3): 217-218.

[3] LIU Da-jie, SHI Wen-zhong, TONG Xiao-hua. Accuracy of Spatial Data and Quality Control Techniques in GIS[M]. Shanghai: Shanghai Science and Technology Literature Publishing House, 1999. (刘大杰, 史文中, 童小华, 等. GIS 空间数据的精度分析与质量控制. 上海: 上海科学技术文献出版社, 1999.)

[4] GABAY Y, DOYTSHER Y. Automatic Adjustment of Line Maps[A]. Proceedings of the GIS/LIS' 94 Annual Convention [C]. Phoenix: [s. n.], 1994, 333-341.

[5] COBB M, CHUNG M, FOLEY H. A Rule-based Approach for the Conflation of Attributed Vector Data[J]. GeoInformatica, 1998, 2(1): 7-35.

[6] WALTER V, FRITSCH D. Matching Spatial Data Sets: a Statical Approach[J]. International Journal of Geographical Information Systems, 1999, 13(5): 445-473.

[7] ZHANG Qiao-ping. Areal Feature Matching and Conflation among Geographic Databases[D]: Wuhan: Wuhan University, 2002. (张桥平. 地图数据库实体匹配与合并技术研究[D]. 武汉: 武汉大学, 2002.)

[8] CINQUE L, LEVIALDI S, MALIZIA A. Shape Description Using Cubic Polynomial Bezier Curves[J]. Pattern Recognition Letters, 1998, 19: 821-828.

[9] WENTZ E A. Shape Analysis in GIS[A]. Proc of ACSM/ASPRS[C]. [s. l.]: [s. n.], 1997. 204-213.

[10] FOLEY H. A Multiple Criteria Based Approach to Performing Conflation in Geographical Information Systems[D]. New Orleans: Tulane University, 1997.

[11] BEERI C, KANZA Y, SAFRA E, SAGIV Y. Object Fusion in Geographic Information Systems[A]. Proceedings of the 30th VLDB Conference[C]. Toronto: [s. n.], 2004: 816-827.

[12] GUO Qing-sheng, DU Xiao-chu, LIU hao. Research on Quantitative Representation and Abstraction of Spatial Topological Relation between Two Regions[J]. Acta Geodaetica et Cartographica Sinica, 2005, 34(2): 123-128. (郭庆胜, 杜晓初, 刘　浩. 空间拓扑关系定量描述与抽象方法研究[J]. 测绘学报, 2005, 34(2): 123-128.)

(责任编辑: 张燕燕)

(上接第 209 页)

系统中的应用对模型的实用性进行了验证。

参考文献:

[1] ZHANG Ze-xun, HUANG Ming-zhi. Temporal GIS Data Structure Discussion[J]. Bulletin of Surveying and Mapping, 1996, 1): 19-22. (张祖勋, 黄明智. 时态 GIS 数据结构的研讨[J]. 测绘通报, 1996, (1): 19-22.)

[2] LIU Ren-yi, LIU Nan. Extension of Spatio-temporal Data Models of Base State with Amendments and Its Implementation in Land Registration Management Systems[J]. Acta Geodaetica et Cartographica Sinica, 2001, 30(2): 168-172. (刘仁义, 刘　南. 基态修正时空数据模型的扩展及在土地产权产籍系统中的实现[J]. 测绘学报, 2001, 30(2): 168-172.)

[3] CHEN Jun, CHEN Shang-chao, TANG Zhi-feng. Representing Temporal Attributes in GISs Using Non-1NF Approach [J]. Journal of Wuhan Technical University of Surveying and Mapping, 1995, 20(1): 12-17. (陈　军, 陈尚超, 唐治锋. 用非第一范式关系表达 GIS 时态属性数据[J]. 武汉测绘科技大学学报, 1995, 20(1): 12-17.)

[4] GONG Jian-ya. An Object-oriented Spatio-temporal Data Model in GIS[J]. Acta Geodaetica et Cartographica Sinica, 1997, 26(4): 289-298. (龚健雅. GIS 中面向对象时空数据模型[J]. 测绘学报, 1997, 26(4): 289-298.)

[5] WORBOYS M F. Object-oriented Approaches to Geo-referenced Information[J]. Geographical Information Systems, 1994, 8(4): 225-245.

[6] CHENG Chang-xiu, ZHOU Cheng-hu, LU Feng. The Improved Base State with Amendments Spatio-temporal Model in the Object-relation GIS[J]. Journal of Image and Graphics, 2003, 8(A)(6): 697-702. (程昌秀, 周成虎, 陆　锋. 对象关系型 GIS 中改进基态修正时空数据模型的实现[J]. 中国图象图形学报, 2003, 8(A)(6): 697-702.)

[7] DAVID S F. Model Driven Architecture: Applying MDA to Enterprise Computing[M]. [s. l.]: Wiley Publishing, 2003.

[8] LU Feng, ZHOU Cheng-hu, WAN Qing. A Feature-based Non-planar Data Model for Urban Traffic Networks[J]. Acta Geodaetica et Cartographica Sinica, 2000, 29(4): 334-341. (陆锋, 周成虎, 万　庆. 基于特征的城市交通网络非平面数据模型[J]. 测绘学报, 2000, 29(4): 334-341.)

[9] SHARMA Chakravarthy, KIM Seung-kyum. Resolution of Time Concepts in Temporal Database[J]. Infomation Science, 1994, 80: 91-125.

[10] PEUQUET D J, DUAN Niu. An Event-based Spatio-temporal Data Model(ESTDM) for Temporal Analysis of Geographical Data[J]. International Journal of Geographical Information Science, 1995, 9 (1): 7-24.

[11] ZHENG Kou-gen, YU Qing-yi, PAN Yun-he. Extension and Implementation of Spatio-temporal Model Based on Event[J], Computer Engineering and Applications, 2003. 3: 45-47. (郑扣根, 余青怡, 潘云鹤. 基于事件对象的时空数据模型的扩展与实现[J]. 计算机工程与应用, 2003. 3: 45-47.)

[12] CAO Zhi-yue, LIU Yue. An Object-oriented Spatio-temporal Data Model[J]. Acta Geodaetica et Cartographica Sinica, 2002, 31(1): 87-92. (曹志月, 刘　岳. 一种面向对象的时空数据模型[J]. 测绘学报, 2002, 31(1): 87-92.)

(责任编辑: 丛树平)